Methodological Summary (Final Exam) - CSC/DSC 462 - FALL 2018

We make use of the following distributions:

$$\begin{aligned} Z &\sim N(0,1) \\ T_\nu &\sim t\text{-distribution with } \nu \text{ degrees of freedom} \\ \chi^2_\nu &\sim \chi^2\text{-distribution with } \nu \text{ degrees of freedom} \\ F_{\nu_1,\nu_2} &\sim F\text{-distribution with } \nu_1 \text{ and } \nu_2 \text{ numerator and denominator degrees of freedom} \end{aligned}$$

Critical values are denoted:

$$\begin{array}{rll} z_\alpha & \text{satisfies} & P(Z > z_\alpha) = \alpha \\ t_{\nu,\alpha} & \text{satisfies} & P(T_\nu > t_{\nu,\alpha}) = \alpha \\ \chi^2_{\nu,\alpha} & \text{satisfies} & P(\chi^2_\nu > \chi^2_{\nu,\alpha}) = \alpha \\ F_{\nu_1,\nu_2,\alpha} & \text{satisfies} & P(F_{\nu_1,\nu_2} > F_{\nu_1,\nu_2,\alpha}) = \alpha \end{array}$$

1 Odds Ratios

We are given an outcome O_+ and two conditions G_1, G_2. The odds ratio for O_+ is defined as

$$OR = \frac{Odds(O_+ \mid G_1)}{Odds(O_+ \mid G_2)} = \frac{P(O_+ \mid G_1)/(1 - P(O_+ \mid G_1))}{P(O_+ \mid G_2)/(1 - P(O_+ \mid G_2))}.$$

Statistically, the OR can be estimated using a 2×2 contingency table:

	O_+	O_-	
G_1	n_{11}	n_{12}	R_1
G_2	n_{21}	n_{22}	R_2
Total	C_1	C_2	n

The formula for the estimate of OR is simply

$$\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Inference proceeds by a natural log transformation $\log(OR)$ of OR, which has standard error

$$SE(\log(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

An approximate $(1-\alpha)100\%$ confidence interval for $\log(OR)$ is therefore

$$CI_{1-\alpha} = \log(\hat{OR}) \pm z_{\alpha/2} SE(\log(OR)).$$

2 Binomial continuity correction

If $X_{bin} \sim bin(n,p)$ and $X_{norm} \sim N(np, np(1-p))$ then we have approximation

$$P(X_{bin} = k) \approx P(k - 0.5 \leq X_{norm} \leq k + 0.5).$$

This means the CDF of X_{bin} should use the approximation

$$F_{X_{bin}}(k) = P(X_{bin} \leq k) \approx P(X_{norm} \leq k + 0.5) = F_{X_{norm}}(k + 0.5).$$

3 Single Population Mean

We assume that there is a population of measurements with a true population mean μ. We take a random sample from this population of size n, and accept the resulting sample mean $\bar{X}_n$ as an estimate of μ. Suppose that the true population variance is σ^2. Then we may define a $(1-\alpha)100\%$ confidence interval to be

$$CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If σ^2 is unknown, then we substitute sample variance S^2 for σ and $t_{n-1,\alpha/2}$ for $z_{\alpha/2}$ giving

$$CI_{1-\alpha} = \bar{X}_n \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

Hypothesis tests use statistics:

$$Z_{obs} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}(\sigma^2 \text{ known}), \;\; T_{obs} = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}}(\sigma^2 \text{ unknown}).$$

1. One sided, lower tailed hypothesis $H_o : \mu \geq \mu_0$ against $H_a : \mu < \mu_0$. If σ^2 known: reject H_o if $Z_{obs} \leq -z_\alpha$, with P-value $\alpha_{obs} = P(Z \leq Z_{obs})$. If σ^2 unknown: reject H_o if $T_{obs} \leq -t_{n-1,\alpha}$, with P-value $\alpha_{obs} = P(T_{n-1} \leq T_{obs})$.

2. One sided, upper tailed hypothesis $H_o : \mu \leq \mu_0$ against $H_a : \mu > \mu_0$. If σ^2 known: reject H_o if $Z_{obs} \geq z_\alpha$, with P-value $\alpha_{obs} = P(Z \geq Z_{obs})$. If σ^2 unknown: reject H_o if $T_{obs} \geq t_{n-1,\alpha}$, with P-value $\alpha_{obs} = P(T_{n-1} \geq T_{obs})$.

3. Two sided hypothesis $H_o : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$. If σ^2 known: reject H_o if $|Z_{obs}| \geq z_{\alpha/2}$, with P-value $\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$. If σ^2 unknown: reject H_o if $|T_{obs}| \geq t_{n-1,\alpha/2}$, with P-value $\alpha_{obs} = 2P(T_{n-1} \leq -|T_{obs}|)$.

4 Difference in Population Means

	Pop'n 1	Pop'n 2
Population mean	μ_1	μ_2
Population variance	σ_1^2	σ_2^2
Sample size	n_1	n_2
Sample mean	$\bar{X}_1$	$\bar{X}_2$
Sample variance	S_1^2	S_2^2
Pooled Variance	$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	

If the samples are paired, $n = n_1 = n_2$, and we calculate paired differences $D_i = X_{2i} - X_{1i}$, $i = 1, \ldots, n$. Then $\bar{D}$ and S_D^2 are the sample mean and sample variance of the differences $D_1, \ldots, D_n$.

The degrees of freedom used in Welch's t-test is given by (round down if needed)

$$\nu_W = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}.$$

Case 1 Samples are independent. Both population variances σ_1^2 and σ_2^2 are known.
Case 2 Samples are independent. The population variances are unknown, but we assume they are equal, so that $\sigma_1^2 = \sigma_2^2$. [Pooled t-test]
Case 3 Samples are independent. The population variances are unknown and we cannot assume that they are equal. [Welch's t-test]
Case 4 Samples are paired. [Paired t-test]

1. One sided, lower tailed hypothesis $H_o : \mu_2 \geq \mu_1$ against $H_a : \mu_2 < \mu_1$. We are looking for evidence that μ_2 is **less than** μ_1.

2. One sided, upper tailed hypothesis $H_o : \mu_2 \leq \mu_1$ against $H_a : \mu_2 > \mu_1$. We are looking for evidence that μ_2 is **greater than** μ_1.

3. Two sided hypothesis $H_o : \mu_2 = \mu_1$ against $H_a : \mu_2 \neq \mu_1$. We are looking for evidence that μ_2 is **not equal to** μ_1.

The following table summarizes, for each case, the procedures for constructing level $1-\alpha$ confidence intervals for $\mu_2 - \mu_1$, and for testing hypotheses (rejection region and P-value α_{obs}).

Case	ν	CI	Test Statistic	Lower Tail	Upper Tail	Two sided
1	-	$\bar{X}_2 - \bar{X}_1 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$Z_{obs} = \frac{\bar{X}_2-\bar{X}_1}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$Z_{obs} \leq -z_\alpha$ $P(Z \leq Z_{obs})$	$Z_{obs} \geq z_\alpha$ $P(Z \geq Z_{obs})$	$\lvert Z_{obs}\rvert \geq z_{\alpha/2}$ $2P(Z \leq -\lvert Z_{obs}\rvert)$
2	$n_1 + n_2 - 2$	$\bar{X}_2 - \bar{X}_1 \pm t_{\nu,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$T_{obs} = \frac{\bar{X}_2-\bar{X}_1}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$T_{obs} \leq -t_{\nu,\alpha}$ $P(T_\nu \leq T_{obs})$	$T_{obs} \geq t_{\nu,\alpha}$ $P(T_\nu \geq T_{obs})$	$\lvert T_{obs}\rvert \geq t_{\nu,\alpha/2}$ $2P(T_\nu \leq -\lvert T_{obs}\rvert)$
3	ν_W	$\bar{X}_2 - \bar{X}_1 \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	$T_{obs} = \frac{\bar{X}_2-\bar{X}_1}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	$T_{obs} \leq -t_{\nu,\alpha}$ $P(T_\nu \leq T_{obs})$	$T_{obs} \geq t_{\nu,\alpha}$ $P(T_\nu \geq T_{obs})$	$\lvert T_{obs}\rvert \geq t_{\nu,\alpha/2}$ $2P(T_\nu \leq -\lvert T_{obs}\rvert)$
4	$n-1$	$\bar{D} \pm t_{\nu,\alpha/2}\frac{S_D}{\sqrt{n}}$	$T_{obs} = \frac{\bar{D}}{S_D/\sqrt{n}}$	$T_{obs} \leq -t_{\nu,\alpha}$ $P(T_\nu \leq T_{obs})$	$T_{obs} \geq t_{\nu,\alpha}$ $P(T_\nu \geq T_{obs})$	$\lvert T_{obs}\rvert \geq t_{\nu,\alpha/2}$ $2P(T_\nu \leq -\lvert T_{obs}\rvert)$

5 Inference for Population Proportions

Suppose the proportion in a population of a certain type is p. To estimate p we take a random sample of size n from the population. If $\hat{p}$ is the proportion in the sample of the type of interest, then this serves as an estimate of p. In particular, $\hat{p} = X/n$ where $X \sim bin(n,p)$. Furthermore, a consequence of the Central Limit Theorem is that, approximately

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

This means that the standard deviation of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. Of course, if we are trying to estimate p that means we don't know its value. We can, however, approximate the standard deviation by substituting $\hat{p}$ for p, to get $\hat{\sigma_{\hat{p}}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. This leads to a level $(1-\alpha)100\%$ confidence interval for p given by

$$CI_{\alpha/2} \quad = \quad \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Suppose we set p_0 to be some hypothetical population proportion, and let p be the true population proportion. We then have the test statistic

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Note that we use the hypothetical value p_0 in the test statistic.

1. One sided, lower tailed hypothesis $H_o : p \geq p_0$ against $H_a : p < p_0$. We are looking for evidence that p is **less than** p_0. Reject H_o if $Z_{obs} \leq -z_\alpha$. $\alpha_{obs} = P(Z \leq Z_{obs})$.

2. One sided, upper tailed hypothesis $H_o : p \leq p_0$ against $H_a : p > p_0$. We are looking for evidence that p is **greater than** p_0. Reject H_o if $Z_{obs} \geq z_\alpha$. $\alpha_{obs} = P(Z \geq Z_{obs})$.

3. Two sided hypothesis $H_o : p = p_0$ against $H_a : p \neq p_0$. We are looking for evidence that p is **not equal to** p_0. Reject H_o if $|Z_{obs}| \geq z_{\alpha/2}$. $\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$.

For smaller sample sizes we can employ for use in $P(Z \leq Z_{obs})$ the continuity correction:

$$Z_{obs} = \frac{X + 0.5 - np_0}{\sqrt{np_0(1-p_0)}}.$$

6 Inference for two population proportions

Assume that the population proportions of interest are p_1 and p_2, and that random samples of size n_1 and n_2 are selected from each. Furthermore, suppose that the proportions of interest observed in the two samples are $\hat{p}_1$ and $\hat{p}_2$. We further assume that the two samples were collected independently. Then a level $(1-\alpha)100\%$ confidence interval for $p_2 - p_1$ is given by

$$CI_{1-\alpha} = \hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Note that for the null hypothesis we may set $p_0 = p_1 = p_2$. Since we usually construct the test statistic to have a certain distribution assuming that H_o is true, it makes sense to combine the two samples to form a single estimate of p_0, referred to as a **pooled** estimate of p_0. This is given by

$$\hat{p}_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

We then have the test statistic

$$Z_{obs} = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Note that we use the pooled estimate $\hat{p}_0$ in the test statistic.

1. One sided, lower tailed hypothesis $H_o : p_2 \geq p_1$ against $H_a : p_2 < p_1$. We are looking for evidence that p_2 is **less than** p_1. Reject H_o if $Z_{obs} \leq -z_\alpha$. $\alpha_{obs} = P(Z \leq Z_{obs})$.

2. One sided, upper tailed hypothesis $H_o : p_2 \leq p_1$ against $H_a : p_2 > p_1$. We are looking for evidence that p_2 is **greater than** p_1. Reject H_o if $Z_{obs} \geq z_\alpha$. $\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$.

3. Two sided hypothesis $H_o : p_2 = p_1$ against $H_a : p_2 \neq p_1$. We are looking for evidence that p_2 is **not equal to** p_1. Reject H_o if $|Z_{obs}| \geq z_{\alpha/2}$. $\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$.

7 Sample Size Estimates

Population Mean. Recall that the confidence interval for a population mean, given population variance σ^2 is $CI_{1-\alpha} = \bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. The margin of error is $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Now suppose before we collect the sample we decide that the margin of error should be E_o, and the confidence level should be $(1-\alpha)100\%$. We can use the previous expression to determine what the sample size should be, giving $n = \left(z_{\alpha/2}\frac{\sigma}{E_o}\right)^2$ as the required sample size. As a technical note, the n calculated will not generally be an integer. In this case we would always round up, as opposed to rounding to the nearest integer. This way, we ensure that the confidence level reported is not overestimated.

Of course, we would rarely know the actual value σ^2, so we would have to substitute an estimate $\hat{\sigma}^2$, giving $n \approx \left(z_{\alpha/2}\frac{\hat{\sigma}}{E_o}\right)^2$. If we had a previous study, or some other prior knowledge, we could rely on that for $\hat{\sigma}^2$. Failing that, a reasonable alternative is to do an initial **pilot study**. This would be a small sample whose primary purpose would be to obtain an estimate of $\hat{\sigma}^2$. An estimate of the required sample size for a fixed margin of error could then be obtained, and the sample then completed.

Population Proportion. The appropriate sample size required to estimate a population proportion can also be estimated using similar reasoning, except that there are some technical differences. Recall that the level $(1-\alpha)100\%$ confidence interval is given by $CI_{\alpha/2} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ so that the margin of error is $E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. If this expression is rearranged we get $n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2$.

A conservative estimate of n is obtained by replacing $\hat{p}$ with 1/2. If it is anticipated that $\hat{p}$ will be much smaller than 1/2 substitute a plausible upper bound p^* for $\hat{p}$.

Two Population Means. Recall that the confidence interval for a difference in population means, given population variances σ_1^2 and σ_2^2, and sample sizes n_1 and n_2 is $CI_{1-\alpha} = \bar{X}_2 - \bar{X}_1 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. The margin of error is therefore $E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. We can, in principle, predict E for any configuration $(\sigma_1^2, \sigma_2^2, n_1, n_2)$. We may specify that $n = n_1 = n_2$, which is referred to as a *balanced design*. In this case we can obtain the formula $n \approx \left(\frac{z_{\alpha/2}}{E}\right)^2(\sigma_1^2 + \sigma_2^2)$.

8 Inference for Variances

If we are given a sample $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$, a level $(1-\alpha)$ confidence interval for σ^2 is given by

$$\frac{S^2}{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)} < \sigma^2 < \frac{S^2}{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}.$$

and a level $(1-\alpha)$ upper confidence bound is given by

$$\sigma^2 < \frac{S^2}{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}$$

where S^2 is the sample variance.

Suppose we have two independent normally distributed populations, with sample sizes n_1, n_2 and sample variances S_1^2, S_2^2. Use test statistic

$$F_{obs} = S_1^2/S_2^2.$$

1. One sided, lower tailed hypothesis $H_o : \sigma_1^2 \geq \sigma_2^2$ against $H_a : \sigma_1^2 < \sigma_2^2$. We are looking for evidence that σ_1^2 is **less than** σ_2^2. Reject H_o if $F_{obs} \leq F_{n_1-1,n_2-1,1-\alpha}$. $\alpha_{obs} = P(F_{n_1-1,n_2-1} \leq F_{obs})$.

2. One sided, upper tailed hypothesis $H_o : \sigma_1^2 \leq \sigma_2^2$ against $H_a : \sigma_1^2 > \sigma_2^2$. We are looking for evidence that σ_1^2 is **greater than** σ_2^2. Reject H_o if $F_{obs} \geq F_{n_1-1,n_2-1,\alpha}$. $\alpha_{obs} = P(F_{n_1-1,n_2-1} \geq F_{obs})$.

3. Two sided hypothesis $H_o : \sigma_2^2 = \sigma_1^2$ against $H_a : \sigma_2^2 \neq \sigma_1^2$. We are looking for evidence that σ_1^2 is **not equal to** σ_2^2. Reject H_o if $F_{obs} \leq F_{n_1-1,n_2-1,1-\alpha/2}$ or $F_{obs} \geq F_{n_1-1,n_2-1,\alpha/2}$. $\alpha_{obs} = 2\min(P(F_{n_1-1,n_2-1} \leq F_{obs}), P(F_{n_1-1,n_2-1} \geq F_{obs}))$.

Note that lower tailed critical values can be obtained from the formula

$$F_{\nu_1,\nu_2,1-\alpha} = 1/F_{\nu_2,\nu_1,\alpha}.$$

9 Goodness of Fit Tests

Suppose we collect a sample of size n of categorical data, consisting of r categories. This will sometimes be numerical data reduced to categories, as is done when constructing a histogram. Suppose we hypothesize that these categories exists in the population according to frequencies $p_1, \ldots, p_r$, and so we wish to test the hypotheses

$$\begin{aligned} H_o &: p_1, \ldots, p_r \text{ are the population frequencies for categories } 1, \ldots, r \\ H_a &: \text{ At least one of the hypothesical frequencies is incorrect} \end{aligned}$$

The statistic we use is given by

$$\begin{aligned} X^2 &= \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}, \text{ where} \\ O_i &= \text{Observed count for category } i = n_i \\ E_i &= \text{Expected count for category } i = np_i \end{aligned}$$

The observed level of significance is $\alpha_{obs} = P(\chi^2_\nu > X^2)$. A size α rejection region is given by $X^2 \geq \chi^2_{r-1,\alpha}$. When small cell sizes are present (≤ 5) Yate's correction may be used in place of X^2:

$$X^2_{Yates} = \sum_{i=1}^{r} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

10 Test for Independence in Contingency Tables

Suppose we are given a contingency table with n_r rows and n_c columns. Let

$$\begin{aligned} r_i &= P(\text{ } i\text{th row event occurs }) \\ c_j &= P(\text{ } j\text{th column event occurs }) \\ p_{ij} &= P(\text{ } i\text{th row event AND } j\text{th column event occur }) \end{aligned}$$

If the ith row event and the jth column event are independent then $p_{ij} = r_i c_j$ so that the hypothesis of row/column independence can be written

$$H_o : p_{ij} = r_i c_j \text{ for all row events } i \text{ and column events } j$$
$$H_a : p_{ij} \neq r_i c_j \text{ for some row event } i \text{ and column event } j.$$

The statistic we use is given by

$$\begin{aligned} X^2 &= \sum_{i}^{n_r} \sum_{j}^{n_c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \text{ where} \\ O_{ij} &= \text{Observed count in cell } i, j \\ E_{ij} &= \text{Expected count in cell } i, j \text{ under } H_o. \end{aligned}$$

The quantity O_{ij} is simply the count given in the cell given by row i and column j. The expected count is $E_{ij} = N r_i c_j$, where N is the total cell count. If we let

$$R_i = \text{Total counts in row } i \quad \text{and} \quad C_j = \text{Total counts in column } j$$

then as an approximation we have

$$r_i \approx \frac{R_i}{N}, \;\; c_j \approx \frac{C_j}{N}, \;\; E_{ij} \approx N \frac{R_i}{N} \frac{C_j}{N} = \frac{R_i C_j}{N}.$$

This means the observed level of significance for rejecting H_o is $\alpha_{obs} = P(\chi^2_\nu \geq X^2)$, where χ^2_ν is a random variable with a χ^2 distribution with $\nu = (n_r - 1)(n_c - 1)$ degrees of freedom. A size α rejection region is given by $X^2 \geq \chi^2_{(n_r-1)(n_c-1),\alpha}$. When small cell sizes are present (≤ 5) Yate's correction may be used in place of X^2:

$$X^2_{Yates} = \sum_{i}^{n_r} \sum_{j}^{n_c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}.$$

11 Sign Test for Paired Comparisons

We are given a paired sample $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, with differences $D_i = X_i - Y_i$, $i = 1, \ldots, n$. The differences can be represented as a random quantity D, and we have null hypothesis H_o : median of D = 0. We evaluate T_+ = number of positive D_i's and T_- = number of negative D_i's. If any of the differences D_i are zero, we discard them and adjust sample size n accordingly.

1. One-sided lower tailed hypothesis H_o : median of D = 0 against H_a : median of D < 0 use $\alpha_{obs} = P(X \leq T_+)$ where $X \sim bin(n, 1/2)$.

2. One-sided upper tailed hypothesis H_o : median of D = 0 against H_a : median of D > 0 use $\alpha_{obs} = P(X \leq T_-)$ where $X \sim bin(n, 1/2)$.

3. Two-sided hypothesis H_o : median of D = 0 against H_a : median of D $\neq$ 0 use $\alpha_{obs} = 2P(X \leq \min(T_-, T_+))$ where $X \sim bin(n, 1/2)$.

12 Wilcoxon Signed Rank Test for Paired Comparisons

We are given a paired sample $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, with differences $D_i = X_i - Y_i$, $i = 1, \ldots, n$. The differences can be represented as a random quantity D, and we have null hypothesis H_o : median of D = 0. For each value we calculate $|D_i|$, and then rank these values (use average ranks for ties). As in the sign test, we assign '+' or '-' according to whether the value is above or below the hypothetical media. We can designate each rank as *positive* or *negative* according to it's sign. We then set T_+ = sum of positive ranks and T_- = sum of negative ranks, with $T_{obs} = \min\{T_-, T_+\}$. If any of the differences D_i are zero, we discard them and adjust sample size n accordingly. For large samples (say, $n > 12$) we can use z-score

$$Z_{obs} = \frac{T_{obs} - \mu_T}{\sigma_T}$$

where $\mu_T = n(n+1)/4$, and $\sigma_T = \sqrt{n(n+1)(2n+1)/24}$, and use approximation $Z_{obs} \sim N(0, 1)$.

1. One-sided lower tailed hypothesis H_o : median of D = 0 against H_a : median of D < 0 reject H_o if $T_{obs} = T_+$ with $\alpha_{obs} = P(Z \leq Z_{obs})$.

2. One-sided upper tailed hypothesis H_o : median of D = 0 against H_a : median of D > 0 reject H_o if $T_{obs} = T_-$ with $\alpha_{obs} = P(Z \leq Z_{obs})$.

3. Two-sided hypothesis H_o : median of D = 0 against H_a : median of D $\neq$ 0 use $\alpha_{obs} = 2P(Z \leq Z_{obs})$.

13 Wilcoxon Rank Sum Test for Independent Samples

We are given independent samples $X_1, \ldots, X_{n_1}$, $Y_1, \ldots, Y_{n_2}$, and we have null hypothesis $H_o : \tilde{\mu}_1 = \tilde{\mu}_2$ where $\tilde{\mu}_1, \tilde{\mu}_2$ are the population medians. Both samples are pooled, then ranked (use average ranks for ties). Set T_1 = sum of ranks from sample 1, T_2 = sum of ranks from sample 2. Usually, samples are labeled (or can be relabeled) so that $n_1 \leq n_2$, and we only need to calculate T_1. Set

$$\mu_1 = n_1(n_1 + n_2 + 1)/2, \ \ \mu_2 = n_2(n_1 + n_2 + 1)/2, \ \text{ and } \ \sigma_W^2 = n_1 n_2 (n_1 + n_2 + 1)/12.$$

$$\text{and } Z_{obs} = \frac{T_1 - \mu_1}{\sigma_W}.$$

1. One-sided lower tailed hypothesis $H_o : \tilde{\mu}_1 \geq \tilde{\mu}_2$ against $H_a : \tilde{\mu}_1 < \tilde{\mu}_2$ use observed significance level $\alpha_{obs} = P(Z \leq Z_{obs})$.

2. One-sided upper tailed hypothesis $H_o : \tilde{\mu}_1 \leq \tilde{\mu}_2$ against $H_a : \tilde{\mu}_1 > \tilde{\mu}_2$ use observed significance level $\alpha_{obs} = P(Z \geq Z_{obs})$.

3. Two-sided hypothesis $H_o : \tilde{\mu}_1 = \tilde{\mu}_2$ against $H_a : \tilde{\mu}_1 \neq \tilde{\mu}_2$ use observed significance level $\alpha_{obs} = 2P(Z \leq -|Z_{obs}|)$.